SLC Walkthrough for Practice Set - Week 1

□ Import required libraries.
□ Load data set.

Let's begin with some hands-on practice exercises:

1. Is the target variable imbalanced?
   HINT:
   Plot the count plot for the target variable, And see, Is the target variable imbalanced or not?

2. Build a full logistic model and calculate the odds for each variable. HINT:
   - Prepare the variables before building the model.
     Ex:
     - □ Scale the numerical variables and encode the categorical variables.
     - □ Split the dataset into train set and test set.
   - Fit logit model.
   - Get params, Params returns the coefficients of all the independent variables.
   - take the exponential of the coefficient of a variable to calculate the odds.

3. Calculate the Specificity and Sensitivity from the confusion matrix of the full model (consider the probability threshold as 0.25).
   HINT:

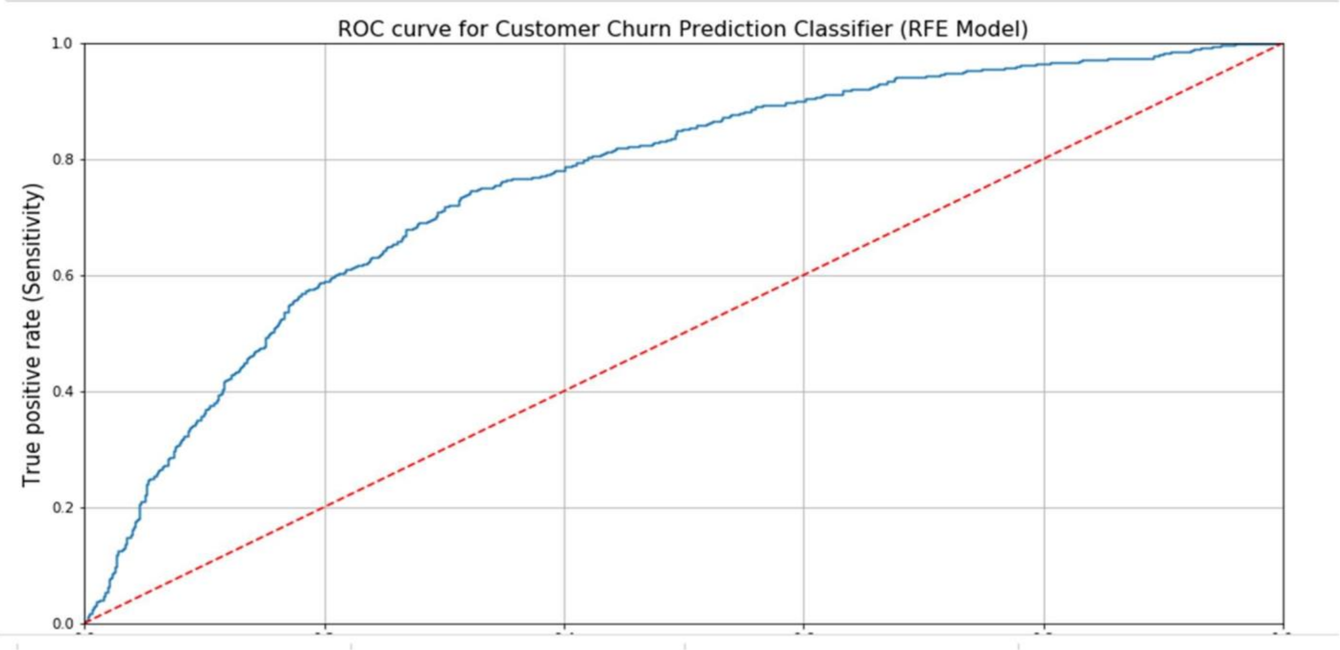
   - Get the probability values from the trained values.
   - And set the threshold as 0.25 and convert probability values into labels by below condition.
     Ex: if probability_value > 0.25 then 1 else 0.

- Use confusion matrix function from sklearn.metrics and calculate specificity and sensitivity.

4. Build a logistic model on the 6 features obtained by RFE and plot the ROC curve

   HINT:

   - Build the RFE feature selection model with number of features to select is 6.
   - Build the Logistic model on the 6 selected features.
   - Use ROC curve to plot scores for different thresholds. Ex: take FPR and TPR to plot the roc curve.



5. Obtain the optimal value of cut-off probability for the full model using the Youden's index.

   HINT:

   - Use roc curve method to get FPR, TPR and Threshold.
   - Create a data frame with 3 columns namely FPR, TPR and Threshold.
   - Take the difference between TPR and FPR and add the difference values as a column in the data frame.

Ex: df_name[“Difference”] = df_name.TPR – df_name.FPR

□ Choose the threshold that has high difference value.

6. Consider the cut-off probability obtained from Youden's index for the full model and calculate the following measures. a. f-1 score.

   b. Accuracy.

   c. Kappa Score.

   HINT:

   - Use probability cut off(threshold) value which you obtained from the question 5.
   - Convert the probability values into class lables namely 0 and 1.

     Ex: if probability value is greater than cut-off value then 1 else 0.
   - Use sklearn metrics to find f1- score and accuracy and kappa score.

7. Identify the variables involved in multicollinearity. HINT:

   Use VIF method to identify the variable.

8. Build a logistic regression model using the categorical variables and the variables obtained after calculating VIF. Also, plot the ROC curve and compute the AUC score (consider the cut-off probability as 0.6).

   HINT:

   - Select the features which are not correlated, we found in the VIF calculation. And also use categorical variable as well.
   - Build the logit model and consider 0.6 as threshold to separate the class labels.
   - Follow the question number 4 points to plot the ROC and AUC plot.

9. Consider the costs of false negatives and false positives as 2 and 0.5 respectively to obtain the optimal cut-off probability for which the total cost will be minimum.
   HINT:
   Multiply False negative with 2 and False positives with 0.5 for different thresholds.
   Use loop to iterate the multiple thresholds, choose a threshold which has less cost.

10. Build a full logistic model using the optimal cut-off probability obtained in Q9. Also, plot the confusion matrix and ROC curve along with the AUC score.
    HINT:
    Convert the probability values into class label with a threshold obtained in the question 9.
    Plot the confusion matrix and roc auc score using sklearn library.